Analyse de Concepts Formels, distributivité et modèles de graphes médians pour la phylogénie

Alain Gély*, Miguel Couceiro*, Amedeo Napoli*

*LORIA (CNRS - Inria Nancy Grand Est - Université de Lorraine),
BP 239, 54506 Vandoeuvre-les-Nancy, France
alain.gely, miguel.couceiro, amedeo.napoli@loria.fr

Résumé. La phylogénie est l'étude des relations de parentés entre les êtres vivants. La classification phylogénétique consiste à classer les êtres vivants à partir de données de phylogénie. Traditionnellement, les modèles utilisés pour ce faire sont les arbres phylogénétiques. Ces arbres ne permettent cependant pas de capturer toute la complexité des phénomènes évolutifs. Du fait de cette complexité, plusieurs arbres peuvent convenir. Pour ne pas privilégier de solution particulière, l'utilisation de graphes médians permet d'encoder l'ensemble des arbres dans un graphe particulier, le graphe médian. Les graphes médians ont des liens étroits avec certains types de treillis, une autre structure souvent utilisée en classification. L'Analyse de Concepts Formels (FCA) a fait des treillis de concepts l'objet central d'étude pour des problèmes d'analyse de données. Dans cet article, nous montrons comment utiliser la FCA pour produire des graphes médians, et nous mettons en avant les verrous techniques à franchir.

1 Introduction

La phylogénie est l'étude des relations de parentés entre les êtres vivants. La classification phylogénétique consiste à classer les êtres vivants à partir de données de phylogénie (variations génétiques par exemple). Traditionnellement, les modèles utilisés pour de telles classifications sont les arbres phylogénétiques. Ces derniers ne permettent cependant pas de capturer toute la complexité des phénomènes évolutifs (mutations inverses, transfert horizontal de gènes). Du fait de cette complexité, plusieurs arbres phylogénétiques peuvent convenir pour les mêmes données initiales.

L'utilisation de graphes médians, introduits par Bandelt (Bandelt et Hedlíková (1983); Bandelt et al. (1999)), permet d'encoder la famille de tous les arbres parcimonieux (minimisant le nombre de changements nécessaires pour passer d'une espèce à l'autre) dans un graphe particulier, le graphe médian. Les graphes médians ont des liens étroits avec certains types de treillis, en particulier les treillis distributifs : tout treillis distributif est un graphe médian, et tout graphe médian peut être considéré comme un semi-treillis vérifiant la propriété de médiane (voir ci-après) sur chaque triplet d'éléments.

Les treillis distributifs sont étudiés dans de nombreux domaines, tant pour leur intérêt théorique que pratique, entre autre pour le lien fort entre ordres partiels et treillis distributifs (Birkhoff (1937)). Birkhoff leur consacre un chapitre dans son ouvrage de référence sur les treillis

(Birkhoff (1967)) et son résultat de représentation des treillis distributifs par des ordres partiels sera central dans les travaux présentés ici.

Les treillis distributifs ne sont qu'une classe particulière de treillis. Pour un treillis quelconque, la distributivité des deux opérations $\vee$ et $\wedge$ n'est pas vérifiée. Les treillis en général sont centraux en Analyse de Concept Formels (FCA) (Ganter et Wille (1999); Barbut et Monjardet) pour l'analyse de données. Uta Priss, dans une série de deux publications (Priss (2012, 2013)) étudie les liens entre FCA et graphes médians pour la phylogénie. Ces deux articles restent au niveau conceptuel et abordent peu les détails algorithmiques sous-jacents.

Dans nos travaux (Gély et al. (2018b,a)), nous nous sommes intéressés a cet aspect algorithmique et avons formalisé une approche. Nous faisons ici un point sur la manière d'obtenir un graphe médian en utilisant les outils de l'Analyse de Concepts Formels. La section 2 détaille les différents modèles possibles, la section 3 montre l'algorithme utilisé. Nous concluerons en section 4 en évoquant les travaux en cours et en donnant quelques perspectives.

2 Modèles

2.1 Graphes médians

Un graphe médian est un graphe $G = (V, E)$ tel que pour tout triplet de sommets $x, y, z \in V$, il existe un unique sommet t à l'intersection de tous les plus courts chemins entre chaque paire de sommets. Les arbres sont un cas particulier de graphe médian.

En phylogénie, le graphe de Buneman (Buneman (1971)), qui est le graphe représentant l'ensemble des arbres phylogénétiques parcimonieux (minimisant le nombre de mutations nécessaires pour passer d'un individu à l'autre) est un graphe médian. Les sommets du graphe de Buneman représentent d'une part les espèces à considérer pour la phylogénie, et d'autre part un certain nombre de sommets latents, ajoutés de façon à vérifier la propriété de médiane. Lorsque les espèces sont décrites par un ensemble de caractères (de type booléen "présent/absent", ou bien "muté/non muté"), il y a une arête entre deux espèces lorsqu'elles ne diffèrent que par un caractère.

Notons que si la phylogénie est parfaite (il n'y a pas eu de mutation inverse ou de transfert latéral), le graphe obtenu est alors naturellement un arbre. Si elle n'est pas parfaite, alors plusieurs arbres peuvent convenir. Chacun de ces arbres est un arbre couvrant du graphe médian obtenu.

2.2 Analyse de Concepts Formels

La classification phylogénétique peut souvent se ramener à utiliser des données binaires entre objets (les espèces) et variables (présence/absence d'une mutation). Ainsi, on peut définir un contexte formel $C = (O, A, I)$, avec O l'ensemble des objets (espèces), A l'ensemble des attributs (mutations) et I une relation binaire entre O et A, telle que pour $o \in O$, $a \in A$ $I(o, a)$ (noté oIa) se lit comme "l'objet o possède l'attribut a" (l'espèce o possède la mutation a).

Un treillis $\mathbf{T} = (T, \leq, \vee, \wedge)$ est un ensemble ordonné muni de deux opérateurs $\vee$ (resp. $\wedge$) correspondant à la borne supérieure (resp. inférieure) de deux éléments de T. Par définition, dans un treillis (contrairement à un ordre quelconque), les bornes supérieures et inférieures

existent toujours. On parlera de semi-treillis si l'on se restreint à l'existence d'une seule de ces deux bornes.

A partir du contexte $C = (O, A, I)$ et des connections de Galois rappelées ci-dessous (Def. 1), on peut définir un treillis $\mathcal{B}(C)$, treillis des concepts du contexte C. Les éléments de ce treillis sont les concepts, c'est à dire les ensembles (X, Y), $X \subseteq O$, $Y \subseteq A$ tels que $X' = Y$ et $Y' = X$. On appelle *extent* l'ensemble X et *intent* l'ensemble Y. En particulier, X et Y vérifient $X = X''$ et $Y = Y''$ et sont des ensembles fermés. La relation d'ordre entre concept est une relation d'inclusion entre les extensions des concepts. Pour plus de détails sur l'Analyse de Concepts Formels, le lecteur pourra se reporter à l'ouvrage de base Ganter et Wille (1999)

Définition 1 *Soit $(O, A, \leq)$ un contexte, on peut définir une connections de Galois entre O et A comme suit :*
— $' : 2^O \to 2^A$, $X' = \{a \mid \forall o \in O, oIa\}$
— $' : 2^A \to 2^O$, $Y' = \{o \mid \forall a \in A, oIa\}$

Un concept représente l'ensemble maximal des individus partageant un ensemble maximal d'attributs. L'ajout d'un nouvel attribut à l'intent (resp. d'un nouvel objet à l'extent) va séparer les objets (resp. attributs) en deux parties strictement non vides : les objets (resp. attributs) en relation avec ce nouvel attribut (resp. objets), et les autres.

Un graphe médian est isomorphe à un semi-treillis distributif particulier. Un treillis des concepts $\mathcal{B}(C)$ n'a pas de raison *a priori* d'être distributif. Il faut donc pouvoir transformer un treillis quelconque en un treillis distributif. Aussi, utiliser le formalisme FCA pour la production de graphe médian va nous amener à décrire plus en détail les treillis distributifs, ce qui est fait dans la section suivante.

2.3 Treillis distributifs

Par définition, un treillis distributif est un treillis pour lequel la loi de distributivité s'applique entre $\vee$ et $\wedge$, c'est à dire : $\forall x, y, z \in T : x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$. Birkhoff s'est énormément intéressé aux treillis distributifs dès les années 30 avec un article dont est issu un des résultats utilisé ici (Birkhoff (1933)). On retrouve aussi la plupart des résultats détaillés dans l'ouvrage Caspard et al. (2012)

Il découle de cette définition plusieurs caractérisations, dont l'une établit le lien entre treillis distributif et graphe médian : un treillis $(T, \leq, \vee, \wedge)$ est distributif ssi $\forall x, y, z$, $(x \wedge y) \vee (y \wedge z) \vee (z \wedge x) = (x \vee y) \wedge (y \vee z) \wedge (z \vee x)$

Or, on peut définir une opération de médiane sur un ensemble M comme une fonction :

$$m : M^3 \to M$$

vérifiant

$$m(a, a, b) = a \text{ et } m(m(a, b, c), d, c) = m(a, m(b, c, d), c)$$

Ainsi, $m(a, b, c) = (a \wedge b) \vee (b \wedge c) \vee (c \wedge a)$ définit une opération de médiane sur un treillis distributif et ce résultat est utilisé par Bandelt (Bandelt et al. (1999)) pour rapprocher les treillis distributifs des graphes médians.

Une autre caractérisation des treillis distributifs est qu'ils ne contiennent ni M_3 ni N_5 comme sous-treillis (un sous-treillis T_1 est un sous-ordre stable pour les opérations $\vee$ et $\wedge$, c'est à dire que si $x, y \in T_1$ alors $x \vee y \in T_1$ et $x \wedge y \in T_1$). Sur la figure 1, on trouve le treillis non distributif N_5 et un treillis distributif. Notons que si N_5 est un sous-ordre du treillis de droite, il n'en est pas un sous-treillis. Pour des raisons de place, M_3, composé d'une antichaîne de 3 éléments, d'un plus petit élément $\perp$ et un plus grand élément $\top$, n'est pas représenté.

Théorème de représentation de Birkhoff. Notre algorithme s'appuie sur un des résultats principaux de Birkhoff pour les treillis distributifs, le théorème de représentation. Ce théorème établit que les treillis distributifs sont en bijection avec les treillis des idéaux d'un ensemble ordonné. De plus, étant donné un treillis distributif, on peut facilement retrouver l'ensemble ordonné dont il est le treillis des idéaux. Il s'agit de l'ensemble ordonné induit par les éléments $\vee$-irréductibles du treillis.

Un *idéal* X est un ensemble ordonné tel que si $x \in X$ et $y < x$, alors $y \in X$. Sur la figure 1, l'ensemble ordonné à droite de la figure correspond par exemple aux idéaux suivants $\{\emptyset, \{1\}, \{3\}, \{1,2\}, \{1,3\}, \{1,2,3\}\}$. Les idéaux de cet ordre, ordonnés par inclusion, forment un treillis isomorphe au treillis représenté à droite. Ce treillis est un treillis distributif.

Un élément $\vee$-irréductible d'un treillis est un élément qui n'est pas borne supérieure de deux éléments autre que lui même. Par exemple, l'élément d'étiquette d (Fig. 1 Droite) n'est pas $\vee$-irréductible puisque $d = 1 \vee 3$. La famille des éléments $\vee$-irréductible de ce treillis est $\{1, (2,b), (3,c)\}$. Ordonné par inclusion, l'ensemble ordonné obtenu est isomorphe à l'ensemble ordonné de droite.

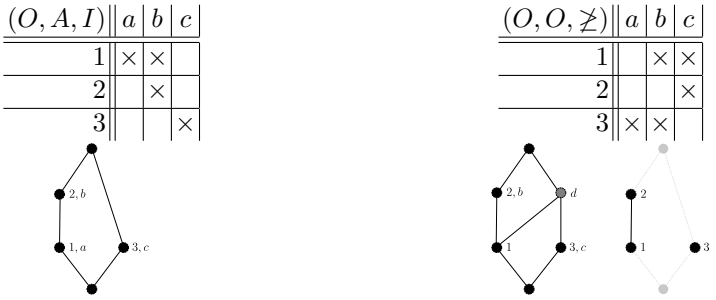

FIG. 1 – *Gauche. Le treillis non distributif N_5 et son contexte. Droite. Un treillis distributif, son contexte et l'ordre induit par les éléments $\vee$-irréductibles. Les deux treillis présentés partagent le même ordre induit par les éléments $\vee$-irréductibles. Le treillis de gauche (N_5) peut se plonger (plongement d'ordre) dans le treillis de droite. Remarquons que les sommets du treillis ont parfois deux étiquettes, selon que l'on considère l'ensemble d'objets ou d'attributs*

3 FCA et construction de graphe médian

Il est possible d'utiliser le théorème de représentation de Birkhoff pour produire un treillis distributif T_d à partir d'un treillis quelconque T tel que T puisse se plonger (plongement

d'ordre) dans T_d. Un treillis distributif étant un graphe médian, c'est donc un moyen de construire un graphe médian en utilisant le formalisme FCA.

On notera que l'entrée des algorithmes est rarement un treillis des concepts, mais plutôt un contexte. Celui-ci contient forcément les éléments $\vee$-irréductibles et il est connu (Ganter et Wille (1999)) qu'à partir du contexte d'un treillis, on peut calculer le contexte du treillis distributif ayant le même ordre induit par les éléments $\vee$-irréductibles. Ce contexte est $C = (O, O, \not\geq)$.

Construction d'un semi-treillis distributif à partir d'un contexte. Si tout treillis distributif est un graphe médian, les graphes médians ne sont pas tous des treillis. Dans ce cas, ils sont isomorphes à un semi-treillis distributif particulier (on étend la notion de distributivité aux semi-treillis en considérant qu'un $\vee$-semi-treillis est distributif si, pour chaque élément minimal o, les éléments qui lui sont supérieurs – le filtre de o – forment un treillis).

Depuis un contexte, on peut construire un semi-treillis en considérant le treillis des concepts privé de son élément minimum. Une méthode pour obtenir un semi-treillis distributif est alors d'appliquer le théorème de représentation de Birkhoff sur chacun des filtres des éléments minimaux. C'est la méthode qui est présentée dans l'algorithme 1. Cette méthode nécessite une boucle externe pour vérifier que les modifications d'un filtre n'ont pas remis en question la distributivité d'un autre filtre. Il est en effet possible que des éléments soient communs à plusieurs filtres.

Algorithme 1 : Construction du contexte du $\vee$-semi-treillis distributif.

```
Données : Un contexte C = (O, A, I)
Résultat : Le contexte C_d = (O, A_d, I_d) d'un semi-treillis distributif
pour chaque o ∈ O, minimal faire
    (P_o, ≤) ← ∅
répéter
    stabilité ← vrai ;
    pour chaque o ∈ O, minimal faire
        calculer P_o l'ensemble ordonné des éléments ∨-irreductibles supérieurs à o
        Produire le contexte C_o = (P_o, P_o, ≱)
        si P_o modifié depuis la dernière itération alors
            stabilité ← faux ;
    Fusionner les différents contextes C_o = (P_o, P_o, ≱)
jusqu'à stabilité
```

4 Conclusion

Il est possible d'utiliser les outils de l'analyse de concepts formels pour produire un graphe médian. Dans les grandes lignes, cela revient à plonger un treillis dans un treillis distributif en utilisant le théorème de représentation de Birkhoff. Cependant, dans le cas où la sortie recherchée est un semi-treillis, il faut considérer les filtres de chaque élément minimal et nous avons

proposé une méthode en ce sens. Parce que les filtres ne sont pas forcément disjoints, l'approche présentée peut produire une solution non minimale (voir Gély et al. (2018a)). D'autre part, des travaux en cours montrent que plusieurs solutions minimales non isomorphes peuvent exister. Il reste maintenant à caractériser une solution minimale canonique et à obtenir un algorithme produisant cette solution.

Références

Bandelt, H.-J., P. Forster, et A. Röhl (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution 16*(1), 37–48.

Bandelt, H.-J. et J. Hedlíková (1983). Median algebras. *Discrete mathematics 45*(1), 1–30.

Barbut, M. et B. Monjardet. Ordre et classification, paris, hachette, 1970. *Zbl0267 6001*.

Birkhoff, G. (1933). On the combination of subalgebras. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 29, pp. 441–464. Cambridge University Press.

Birkhoff, G. (1937). Rings of sets. *Duke Math. J. 3*(3), 443–454.

Birkhoff, G. (1967). *Lattice Theory* (3rd ed.). Providence : American Mathematical Society.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*.

Caspard, N., B. Leclerc, et B. Monjardet (2012). *Finite ordered sets : concepts, results and uses*. Number 144. Cambridge University Press.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer.

Gély, A., M. Couceiro, Y. Namir, et A. Napoli (2018b). Contribution à l'étude de la distributivité d'un treillis de concepts. In *Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018*, pp. 107–118.

Gély, A., M. Couceiro, et A. Napoli (2018a). Steps towards achieving distributivity in formal concept analysis. In *Proceedings of the Fourteenth International Conference on Concept Lattices and Their Applications, CLA 2018, Olomouc, Czech Republic, June 12-14, 2018.*, pp. 105–116.

Priss, U. (2012). Concept lattices and median networks. In *CLA*, pp. 351–354.

Priss, U. (2013). Representing median networks with concept lattices. In *ICCS*, pp. 311–321. Springer.

Summary

Philogenetic classification uses phylogeny data to classify species. The more traditionnal models are phylogenetic trees. Nevertheless, trees miss some complexity of evolution, and so, several trees should be used. Median graphs permit to encode all these trees in a unique structure. Median graphs have links with some kind of lattices, another structure used in data analysis. Concept lattices are the central object of Formal Concept Analysis (FCA), a framework for data analysis. In this article, we show how to use FCA to produce median graphs and we enlight some technical difficulties to be tackled.